PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12Q 1/68, C07H 21/02, 21/04

A1

(11) International Publication Number: WO 95/11995

(43) International Publication Date: 4 May 1995 (04.05.95)

(21) International Application Number: PCT/US94/12305

(22) International Filing Date: 26 October 1994 (26.10.94)

(30) Priority Data:

08/143,312	26 October 1993 (26.10.93)	US
08/284,064	2 August 1994 (02.08.94)	US

(60) Parent Application or Grant
(63) Related by Continuation
US 08/284,064 (CIP)
Filed on 2 August 1994 (02.08.94)

(71) Applicant *(for all designated States except US)*: AFFYMAX TECHNOLOGIES N.V. [NL/NL]; De Ruyderkade 62, Curacao (AN).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: CHEE, Mark [US/US]; 3199 Waverly Street, Palo Alto, CA 94306 (US). CRONIN, Maureen, T. [US/US]; 771 Anderson Drive, Los Altos, CA 94024 (US). FODOR, Stephen, P., A. [US/US]; 3863 Nathan Way, Palo Alto, CA 94303 (US). GINGERAS, Thomas, R. [US/US]; 1568 Vista Club Circle, Santa Clara, CA 95054 (US). HUANG, Xiaohua, C. [-/US]; 937 Jackson Street, Mountain View, CA 94043 (US). HUBBELL, Earl, A. [US/US]; 1929 Crisanto #425, Mountain View, CA 94040 (US). LIPSHUTZ, Robert, J. [US/US]; 970 Palo Alto Avenue, Palo Alto, CA 94301 (US). LOBBAN, Peter, E. [US/US]; 273 Lowell Avenue, Palo Alto, CA 94301 (US). MIYADA, Charles, Garrett [US/US]; Sunnyvale, CA (US). MORRIS, MacDonald, S. [US/US]; P.O. Box 720488, San Jose, CA 95172 (US). SHAH, Nila [IN/US]; 12135 Saraglen, Saratoga, CA 95070 (US). SHELDON, Edward, L. [US/US]; 2031 Ashton Avenue, Menlo Park, CA 94025 (US).

(74) Agents: LIEBESCHUETZ, Joseph et al.; Townsend and Townsend Khourie and Crew, Steuart Street Tower, 20th floor, One Market Plaza, San Francisco, CA 94105 (US).

(81) Designated States: AM, AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, JP, KE, KG, KP, KR, KZ, LK, LR, LT, LU, LV, MD, MG, MN, MW, NL, NO, NZ, PL, PT, RO, RU, SD, SE, SI, SK, TJ, TT, UA, US, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: ARRAYS OF NUCLEIC ACID PROBES ON BIOLOGICAL CHIPS

(57) Abstract

The invention provides chips of immobilized probes, and methods employing the chips, for comparing a reference polynucleotide sequence of known sequence with a target sequence showing substantial similarity with the reference sequence, but differing in the presence of e.g., mutations.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgystan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo			SE	Sweden
CH	Switzerland	KR	Republic of Korea	SI	Slovenia
CI	Côte d'Ivoire	KZ	Kazakhstan	SK	Slovakia
CM	Cameroon	LI	Liechtenstein	SN	Senegal
CN	China	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
CZ	Czech Republic	LV	Latvia	TJ	Tajikistan
DE	Germany	MC	Monaco	TT	Trinidad and Tobago
DK	Denmark	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	US	United States of America
FI	Finland	ML	Mali	UZ	Uzbekistan
FR	France	MN	Mongolia	VN	Viet Nam
GA	Gabon				

ARRAYS OF NUCLEIC ACID PROBES ON BIOLOGICAL CHIPS

Cross-Reference to Related Application

This application is a continuation-in-part of USSN 08/284,064, filed August 2, 1994, which is a continuation-in-part of USSN 08/143,312, filed October 26, 1993, each of which is incorporated by reference in its entirety for all purposes.

Research leading to the invention was funded in part by NIH grant No. 1R01HG00813-01, and the government may have certain rights to the invention.

Background of the Invention

Field of the Invention

The present invention provides arrays of oligonucleotide probes immobilized in microfabricated patterns on silica chips for analyzing molecular interactions of biological interest. The invention therefore relates to diverse fields impacted by the nature of molecular interaction, including chemistry, biology, medicine, and medical diagnostics.

Description of Related Art

Oligonucleotide probes have long been used to detect complementary nucleic acid sequences in a nucleic acid of interest (the "target" nucleic acid). In some assay formats, the oligonucleotide probe is tethered, *i.e.*, by covalent attachment, to a solid support, and arrays of oligonucleotide probes immobilized on solid supports have been used to detect specific nucleic acid sequences in a target nucleic acid. *See, e.g.*, PCT patent publication Nos. WO 89/10977 and 89/11548. Others have proposed the use of large numbers of oligonucleotide probes to provide the complete nucleic acid sequence of a target nucleic acid but failed to provide an enabling method for using arrays of immobilized probes for this purpose. *See* U.S. Patent Nos. 5,202,231 and 5,002,867 and PCT patent publication No. WO 93/17126.

The development of VLSIPS™ technology has provided methods for making very large arrays of oligonucleotide probes in very small arrays. *See* U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated herein by reference. U.S. Patent application Serial No. 082,937, filed June 25, 1993, describes methods for making arrays of oligonucleotide probes that can be used to provide the complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific nucleotide sequence.

Microfabricated arrays of large numbers of oligonucleotide probes, called "DNA chips" offer great promise for a wide variety of applications. New methods and reagents are required to realize this promise, and the present invention helps meet that need.

SUMMARY OF THE INVENTION

The invention provides several strategies employing immobilized arrays of probes for comparing a reference sequence of known sequence with a target sequence showing substantial similarity with the reference sequence, but differing in the presence of, *e.g.*, mutations. In a first embodiment, the invention provides a tiling strategy employing an array of immobilized oligonucleotide probes comprising at least two sets of probes. A first probe set comprises a plurality of probes, each probe comprising a segment of at least three nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence. A second probe set comprises a corresponding probe for each probe in the first probe set, the corresponding probe in the second probe set being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the two corresponding probes from the first and second probe sets. The probes in the first probe set have at

least two interrogation positions corresponding to two contiguous nucleotides in the reference sequence. One interrogation position corresponds to one of the contiguous nucleotides, and the other interrogation position to the other.

In a second embodiment, the invention provides a tiling strategy employing an array comprising four probe sets. A first probe set comprises a plurality of probes, each probe comprising a segment of at least three nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence. Second, third and fourth probe sets each comprise a corresponding probe for each probe in the first probe set. The probes in the second, third and fourth probe sets are identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the four corresponding probes from the four probe sets. The first probe set often has at least 100 interrogation positions corresponding to 100 contiguous nucleotides in the reference sequence. Sometimes the first probe set has an interrogation position corresponding to every nucleotide in the reference sequence. The segment of complementarity within the probe set is usually about 9-21 nucleotides. Although probes may contain leading or trailing sequences in addition to the 9-21 sequences, many probes consist exclusively of a 9-21 segment of complementarity.

In a third embodiment, the invention provides immobilized arrays of probes tiled for multiple reference sequences. One such array comprises at least one pair of first and second probe groups, each group comprising first and second sets of probes as defined in the first embodiment. Each probe in the first probe set from the first group is exactly complementary to a subsequence of a first reference sequence, and each probe in the first probe set from the second group is exactly

complementary to a subsequence of a second reference sequence. Thus, the first group of probes are tiled with respect to a first reference sequence and the second group of probes with respect to a second reference sequence. Each group of probes can also include third and fourth sets of probes as defined in the second embodiment. In some arrays of this type, the second reference sequence is a mutated form of the first reference sequence.

In a fourth embodiment, the invention provides arrays for block tiling. Block tiling is a species of the general tiling strategies described above. The usual unit of a block tiling array is a group of probes comprising a wildtype probe, a first set of three mutant probes and a second set of three mutant probes. The wildtype probe comprises a segment of at least three nucleotides exactly complementary to a subsequence of a reference sequence. The segment has at least first and second interrogation positions corresponding to first and second nucleotides in the reference sequence. The probes in the first set of three mutant probes are each identical to a sequence comprising the wildtype probe or a subsequence of at least three nucleotides thereof including the first and second interrogation positions, except in the first interrogation position, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the second set of three mutant probes are each identical to a sequence comprising the wildtype probes or a subsequence of at least three nucleotides thereof including the first and second interrogation positions, except in the second interrogation position, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe.

In a fifth embodiment, the invention provides methods of comparing a target sequence with a reference sequence using arrays of immobilized pooled probes. The arrays employed in these methods represent a further species of the general tiling arrays noted above. In these methods, variants of a reference sequence differing from the reference sequence in at least one nucleotide are identified and each is assigned a

designation. An array of pooled probes is provided, with each pool occupying a separate cell of the array. Each pool comprises a probe comprising a segment exactly complementary to each variant sequence assigned a particular designation. The array is then contacted with a target sequence comprising a variant of the reference sequence. The relative hybridization intensities of the pools in the array to the target sequence are determined. The identity of the target sequence is deduced from the pattern of hybridization intensities. Often, each variant is assigned a designation having at least one digit and at least one value for the digit. In this case, each pool comprises a probe comprising a segment exactly complementary to each variant sequence assigned a particular value in a particular digit. When variants are assigned successive numbers in a numbering system of base m having n digits, n x (m-1) pooled probes are used are used to assign each variant a designation.

In a sixth embodiment, the invention provides a pooled probe for trellis tiling, a further species of the general tiling strategy. In trellis tiling, the identity of a nucleotide in a target sequence is determined from a comparison of hybridization intensities of three pooled trellis probes. A pooled trellis probe comprises a segment exactly complementary to a subsequence of a reference sequence except at a first interrogation position occupied by a pooled nucleotide N, a second interrogation position occupied by a pooled nucleotide selected from the group of three consisting of (1) M or K, (2) R or Y and (3) S or W, and a third interrogation position occupied by a second pooled nucleotide selected from the group. The pooled nucleotide occupying the second interrogation position comprises a nucleotide complementary to a corresponding nucleotide from the reference sequence when the second pooled probe and reference sequence are maximally aligned, and the pooled nucleotide occupying the third interrogation position comprises a nucleotide complementary to a corresponding nucleotide from the reference sequence when the third pooled probe and the reference

sequence are maximally aligned. Standard IUPAC nomenclature is used for describing pooled nucleotides.

In trellis tiling, an array comprises at least first, second and third cells, respectively occupied by first, second and third pooled probes, each according to the generic description above. However, the segment of complementarity, location of interrogation positions, and selection of pooled nucleotide at each interrogation position may or may not differ between the three pooled probes subject to the following constraint. One of the three interrogation positions in each of the three pooled probes must align with the same corresponding nucleotide in the reference sequence. This interrogation position must be occupied by a N in one of the pooled probes, and a different pooled nucleotide in each of the other two pooled probes.

In a seventh embodiment, the invention provides arrays for bridge tiling. Bridge tiling is a species of the general tiling strategies noted above, in which probes from the first probe set contain more than one segment of complementarity. In bridge tiling, a nucleotide in a reference sequence is usually determined from a comparison of four probes. A first probe comprises at least first and second segments, each of at least three nucleotides and each exactly complementary to first and second subsequences of a reference sequences. The segments including at least one interrogation position corresponding to a nucleotide in the reference sequence. Either (1) the first and second subsequences are noncontiguous in the reference sequence, or (2) the first and second subsequences are contiguous and the first and second segments are inverted relative to the first and second subsequences. The arrays further comprises second, third and fourth probes, which are identical to a sequence comprising the first probe or a subsequence thereof comprising at least three nucleotides from each of the first and second segments, except in the at least one interrogation position, which differs in each of the probes. In a species of bridge tiling, referred to as deletion tiling, the first and second subsequences are separated by one or two nucleotides in the reference sequence.

In an eighth embodiment, the invention provides arrays of probes for multiplex tiling. Multiplex tiling is a strategy, in which the identity of two nucleotides in a target sequence is determined from a comparison of the hybridization intensities of four probes, each having two interrogation positions. Each of the probes comprising a segment of at least 7 nucleotides that is exactly complementary to a subsequence from a reference sequence, except that the segment may or may not be exactly complementary at two interrogation positions. The nucleotides occupying the interrogation positions are selected by the following rules: (1) the first interrogation position is occupied by a different nucleotide in each of the four probes, (2) the second interrogation position is occupied by a different nucleotide in each of the four probes, (3) in first and second probes, the segment is exactly complementary to the subsequence, except at no more than one of the interrogation positions, (4) in third and fourth probes, the segment is exactly complementary to the subsequence, except at both of the interrogation positions.

In a ninth embodiment, the invention provides arrays of immobilized probes including helper mutations. Helper mutations are useful for, *e.g.*, preventing self-annealing of probes having inverted repeats. In this strategy, the identity of a nucleotide in a target sequence is usually determined from a comparison of four probes. A first probe comprises a segment of at least 7 nucleotides exactly complementary to a subsequence of a reference sequence except at one or two positions, the segment including an interrogation position not at the one or two positions. The one or two positions are occupied by helper mutations. Second, third and fourth mutant probes are each identical to a sequence comprising the wildtype probe or a subsequence thereof including the interrogation position and the one or two positions, except in the interrogation position, which is occupied by a different nucleotide in each of the four probes.

In a tenth embodiment, the invention provides arrays of probes comprising at least two probe sets, but lacking a probe set comprising probes that are perfectly matched to a

reference sequence. Such arrays are usually employed in methods in which both reference and target sequence are hybridized to the array. The first probe set comprising a plurality of probes, each probe comprising a segment exactly complementary to a subsequence of at least 3 nucleotides of a reference sequence except at an interrogation position. The second probe set comprises a corresponding probe for each probe in the first probe set, the corresponding probe in the second probe set being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the interrogation position, except that the interrogation position is occupied by a different nucleotide in each of the two corresponding probes and the complement to the reference sequence.

In an eleventh embodiment, the invention provides methods of comparing a target sequence with a reference sequence comprising a predetermined sequence of nucleotides using any of the arrays described above. The methods comprise hybridizing the target nucleic acid to an array and determining which probes, relative to one another, in the array bind specifically to the target nucleic acid. The relative specific binding of the probes indicates whether the target sequence is the same or different from the reference sequence. In some such methods, the target sequence has a substituted nucleotide relative to the reference sequence in at least one undetermined position, and the relative specific binding of the probes indicates the location of the position and the nucleotide occupying the position in the target sequence. In some methods, a second target nucleic acid is also hybridized to the array. The relative specific binding of the probes then indicates both whether the target sequence is the same or different from the reference sequence, and whether the second target sequence is the same or different from the reference sequence. In some methods, when the array comprises two groups of probes tiled for first and second reference sequences, respectively, the relative specific binding of probes in the first group indicates whether the

target sequence is the same or different from the first reference sequence. The relative specific binding of probes in the second group indicates whether the target sequence is the same or different from the second reference sequence. Such methods are particularly useful for analyzing heterologous alleles of a gene. Some methods entail hybridizing both a reference sequence and a target sequence to any of the arrays of probes described above. Comparison of the relative specific binding of the probes to the reference and target sequences indicates whether the target sequence is the same or different from the reference sequence.

In a twelfth embodiment, the invention provides arrays of immobilized probes in which the probes are designed to tile a reference sequence from a human immunodeficiency virus. Reference sequences from either the reverse transcriptase gene or protease gene of HIV are of particular interest. Some chips further comprise arrays of probes tiling a reference sequence from a 16S RNA or DNA encoding the 16S RNA from a pathogenic microorganism. The invention further provides methods of using such arrays in analyzing a HIV target sequence. The methods are particularly useful where the target sequence has a substituted nucleotide relative to the reference sequence in at least one position, the substitution conferring resistance to a drug use in treating a patient infected with a HIV virus. The methods reveal the existence of the substituted nucleotide. The methods are also particularly useful for analyzing a mixture of undetermined proportions of first and second target sequences from different HIV variants. The relative specific binding of probes indicates the proportions of the first and second target sequences.

In a thirteenth embodiment, the invention provides arrays of probes tiled based on reference sequence from a CFTR gene. A preferred array comprises at least a group of probes comprising a wildtype probe, and five sets of three mutant probes. The wildtype probe is exactly complementary to a subsequence of a reference sequence from a cystic fibrosis gene, the segment having at least five interrogation positions

corresponding to five contiguous nucleotides in the reference sequence. The probes in the first set of three mutant probes are each identical to the wildtype probe, except in a first of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the second set of three mutant probes are each identical to the wildtype probe, except in a second of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the third set of three mutant probes are each identical to the wildtype probe, except in a third of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the fourth set of three mutant probes are each identical to the wildtype probe, except in a fourth of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the fifth set of three mutant probes are each identical to the wildtype probe, except in a fifth of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. Preferably, a chip comprises two such groups of probes. The first group comprises a wildtype probe exactly complementary to a first reference sequence, and the second group comprises a wildtype probe exactly complementary to a second reference sequence that is a mutated form of the first reference sequence.

The invention further provides methods of using the arrays of the invention for analyzing target sequences from a CFTR gene. The methods are capable of simultaneously analyzing first and second target sequences representing heterozygous alleles of a CFTR gene.

In a fourteenth embodiment, the invention provides arrays of probes tiling a reference sequence from a p53 gene, an hMLH1 gene and/or an MSH2 gene. The invention further provides methods of using the arrays described above to

analyze these genes. The method are useful, *e.g.*, for diagnosing patients susceptible to developing cancer.

In a fifteenth embodiment, the invention provides arrays of probes tiling a reference sequence from a mitochondrial genome. The reference sequence may comprise part or all of the D-loop region, or all, or substantially all, of the mitochondrial genome. The invention further provides method of using the arrays described above to analyze target sequences from a mitochondrial genome. The methods are useful for identifying mutations associated with disease, and for forensic, epidemiological and evolutionary studies.

BRIEF DESCRIPTION OF THE FIGURES

Fig. 1: Basic tiling strategy. The figure illustrates the relationship between an interrogation position (I) and a corresponding nucleotide (n) in the reference sequence, and between a probe from the first probe set and corresponding probes from second, third and fourth probe sets.

Fig. 2: Segment of complementarity in a probe from the first probe set.

Fig. 3: Incremental succession of probes in a basic tiling strategy. The figure shows four probe sets, each having three probes. Note that each probe differs from its predecessor in the same set by the acquisition of a 5' nucleotide and the loss of a 3' nucleotide, as well as in the nucleotide occupying the interrogation position.

Fig. 4: Exemplary arrangement of lanes on a chip. The chip shows four probe sets, each having five probes and each having a total of five interrogation positions (I1-I5), one per probe.

Fig. 5: Hybridization pattern of chip having probes laid down in lanes. Dark patches indicate hybridization. The probes in the lower part of the figure occur at the column of the array indicated by the arrow when the probes length is 15 and the interrogation position 7.

Fig. 6: Strategies for detecting deletion and insertion mutations. Bases in brackets may or may not be present.

Fig. 7: Block tiling strategy. The probe from the first probe set has three interrogation positions. The probes from the other probe sets have only one of these interrogation positions.

Fig. 8: Multiplex tiling strategy. Each probe has two interrogation positions.

Fig. 9. Helper mutation strategy. The segment of complementarity differs from the complement of the reference sequence at a helper mutation as well as the interrogation position.

Fig. 10 Layout of probes on the HV 407 chip. The figure shows successive rows of sequence each of which is subdivided into four lanes. The four lanes correspond to the A-, C-, G- and T-lanes on the chip. Each probe is represented by the nucleotide occupying its interrogation position. The letter "N" indicates a control probe or empty column. The different sized-probes are laid out in parallel. That is, from top-to-bottom, a row of 13 mers is followed by a row of 15 mers, which is followed by a row of 17 mers, which is followed by a row of 19 mers.

Fig. 11 Fluorescence pattern of HV 407 hybridized to a target sequence (pPol19) identical to the chips reference sequence.

Fig. 12 Sequence read from HV 407 chip hybridized to pPol19 and 4MUT18 (separate experiments). The reference sequence is designated "wildtype." Beneath the reference sequence are four rows of sequence read from the chip hybridized to the pPol19 target, the first row being read from 13 mers, the second row from 15 mers, the third row from 17 mers and the fourth row from 19 mers. Beneath these sequences, there are four further rows of sequence read from the chip hybridized to the HXB2 target. Successive rows are read from 13 mers, 15 mers, 17 mers and 19 mers. Each nucleotide in a row is called from the relative fluorescence intensities of probes in A-, C-, G- and T-lanes. Regions of ambiguous sequence read from the chip are highlighted. The strain differences between the HBX2 sequence and the reference sequence that were correctly detected are indicated (*), and

those that could not be called are indicated (o). (The nucleotide at position 417 was read correctly in some experiments). The location of some mutations known to be associated with drug resistance that occur in readable regions of the chip are shown above (codon number) and below (mutant nucleotide) the sequence designated "wildtype." The locations of primer used to amplify the target sequence are indicated by arrows.

Fig. 13: Detection of mixed target sequences. The mutant target differs from the wildtype by a single mutation in codon 67 of the reverse transcriptase gene. Each different sized group of probes has a column of four probes for reading the nucleotide in which the mutation occurs. The four probes occupying a column are represented by a single probe in the figure with the symbol (o) indicating the interrogation position, which is occupied by a different nucleotide in each probe.

Fig. 14: Fluorescence intensities of target bound to 13 mers and 15 mers for different proportions of mutant and wildtype target. The fluorescence intensities are from probes having interrogation positions for reading the nucleotide at which the mutant and wildtype targets diverge.

Fig. 15: Sequence read from protease chip from four clinical samples before and after treatment with ddI>.

Fig. 16: Block tiling array of probes for analyzing a CFTR point mutation. Each probe show actually represents four probes, with one probe having each of A, C, G or T at the interrogation position N. In the order shown, the first probe shown on the left is tiled from the wildtype reference sequence, the second probe from the mutant sequence, and so on in alternating fashion. Note that all of the probes are identical except at the interrogation position, which shifts one position between successive probes tiled from the same reference sequence (*e.g.*, the first, third and fifth probes in the left hand column.) The grid shows the hybridization intensities when the array is hybridized to the reference sequence.

Fig. 17: Hybridization pattern for heterozygous target. The figure shows the hybridization pattern when the array of the previous figure is hybridized to a mixture of mutant and wildtype reference sequences.

Fig. 18, in panels A, B, and C, shows an image made from the region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to a wild-type target; in panel C, the chip was hybridized to a mutant ΔF508 target; and in panel B, the chip was hybridized to a mixture of the wild-type and mutant targets.

Fig. 19, in sheets 1 - 3, corresponding to panels A, B, and C of Fig. 18, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

Fig. 20, in panels A, B, and C, shows an image made from a region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to the wt480 target; in panel C, the chip was hybridized to the mu480 target; and in panel B, the chip was hybridized to a mixture of the wild-type and mutant targets.

Fig. 21, in sheets 1 - 3, corresponding to panels A, B, and C of Fig. 20, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

Fig. 22, in panels A and B, shows an image made from a region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to nucleic acid derived from the genomic DNA of an individual with wild-type ΔF508 sequences; in panel B, the target nucleic acid originated from a heterozygous (with respect to the ΔF508 mutation) individual.

Fig. 23, in sheets 1 and 2, corresponding to panels A and B of Fig. 22, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

Fig. 24: Hybridization of homozygous wildtype (A) and heterozygous (B) target sequences from exon 11 of the CFTR gene to a block tiling array designed to detect G551D and Q552X mutations in CFTR gene.

Fig. 25: Hybridization of homozygous wildtype (A) and ΔF508 mutant (B) target sequences from exon 10 of the CFTR gene to a block tiling array designed to detect mutations, ΔF508, ΔI507 and F508C.

Fig. 26: Hybridization of heterozygous mutant target sequences, ΔF508/F508C, to the array of Fig. 25.

Fig. 27 shows the alignment of some of the probes on a p53 DNA chip with a 12-mer model target nucleic acid.

Fig. 28 shows a set of 10-mer probes for a p53 exon 6 DNA chip.

Fig. 29 shows that very distinct patterns are observed after hybridization of p53 DNA chips with targets having different 1 base substitutions. In the first image in Fig. 29, the 12-mer probes that form perfect matches with the wild-type target are in the first row (top). The 12-mer probes with single base mismatches are located in the second, third, and fourth rows and have much lower signals.

Fig. 30, in graphs 2, 3, and 4, graphically depicts the data in Fig. 29. On each graph, the X ordinate is the position of the probe in its row on the chip, and the Y ordinate is the signal at that probe site after hybridization.

Fig. 31 shows the results of hybridizing mixed target populations of WT and mutant p53 genes to the p53 DNA chip.

Fig. 32, in graphs 1-4, shows (see Fig. 30 as well) the hybridization efficiency of a 10-mer probe array as compared to a 12-mer probe array.

Fig. 33 shows an image of a p53 DNA chip hybridized to a target DNA.

Fig. 34 illustrates how the actual sequence was read from the chip shown in Fig. 33. Gaps in the sequence of letters in the WT rows correspond to control probes or sites. Positions at which bases are miscalled are represented by letters in italic type in cells corresponding to probes in which the WT bases have been substituted by other bases.

Fig. 35 shows the human mitochondrial genome; "O_H" is the H strand origin of replication, and arrows indicate the cloned unshaded sequence.

Fig. 36 shows the image observed from application of a sample of mitochondrial DNA derived nucleic acid (from the mt4 sample) on a DNA chip.

Fig. 37 is similar to Fig. 36 but shows the image observed from the mt5 sample.

Fig. 38 shows the predicted difference image between the mt4 and mt5 samples on the DNA chip based on mismatches between the two samples and the reference sequence.

Fig. 39 shows the actual difference image observed for the mt4 and mt5 samples.

Fig. 40, in sheets 1 and 2, shows a plot of normalized intensities across rows 10 and 11 of the array and a tabulation of the mutations detected.

Fig. 41 shows the discrimination between wild-type and mutant hybrids obtained with the chip. A median of the six normalized hybridization scores for each probe was taken; the graph plots the ratio of the median score to the normalized

hybridization score versus mean counts. A ratio of 1.6 and mean counts above 50 yield no false positives.

Fig. 42 illustrates how the identity of the base mismatch may influence the ability to discriminate mutant and wild-type sequences more than the position of the mismatch within an oligonucleotide probe. The mismatch position is expressed as % of probe length from the 3'-end. The base change is indicated on the graph.

Fig. 43 provides a 5' to 3' sequence listing of one target corresponding to the probes on the chip. X is a control probe. Positions that differ in the target (*i.e.*, are mismatched with the probe at the designated site) are in bold.

Fig. 44 shows the fluorescence image produced by scanning the chip described in Fig. 17 when hybridized to a sample.

Fig. 45 illustrates the detection of 4 transitions in the target sequence relative to the wild-type probes on the chip in Fig. 44.

Fig. 46: VLSIPS™ technology applied to the light directed synthesis of oligonucleotides. Light (hv) is shone through a mask (M_1) to activate functional groups (-OH) on a surface by removal of a protecting group (X). Nucleoside building blocks protected with photoremovable protecting groups (T-X, C-X) are coupled to the activated areas. By repeating the irradiation and coupling steps, very complex arrays of oligonucleotides can be prepared.

Fig. 47: Use of the VLSIPS™ process to prepare "nucleoside combinatorials" or oligonucleotides synthesized by coupling all four nucleosides to form dimers, trimers, and so forth.

Fig. 48: Deprotection, coupling, and oxidation steps of a solid phase DNA synthesis method.

Fig. 49: An illustrative synthesis route for the nucleoside building blocks used in the VLSIPS™ method.

Fig. 50: A preferred photoremovable protecting group, MeNPOC, and preparation of the group in active form.

Fig. 51: Detection system for scanning a DNA chip.

DETAILED DESCRIPTION OF THE INVENTION

The invention provides a number of strategies for comparing a polynucleotide of known sequence (a reference sequence) with variants of that sequence (target sequences). The comparison can be performed at the level of entire genomes, chromosomes, genes, exons or introns, or can focus on individual mutant sites and immediately adjacent bases. The strategies allow detection of variations, such as mutations or polymorphisms, in the target sequence irrespective whether a particular variant has previously been characterized. The strategies both define the nature of a variant and identify its location in a target sequence.

The strategies employ arrays of oligonucleotide probes immobilized to a solid support. Target sequences are analyzed by determining the extent of hybridization at particular probes in the array. The strategy in selection of probes facilitates distinction between perfectly matched probes and probes showing single-base or other degrees of mismatches. The strategy usually entails sampling each nucleotide of interest in a target sequence several times, thereby achieving a high degree of confidence in its identity. This level of confidence is further increased by sampling of adjacent nucleotides in the target sequence to nucleotides of interest. The number of probes on the chip can be quite large (*e.g.*, 10^5-10^6). However, usually only a small proportion of the total number of probes of a given length are represented. Some advantage of the use of only a small proportion of all possible probes of a given length include: (i) each position in the array is highly informative, whether or not hybridization occurs; (ii) nonspecific hybridization is minimized; (iii) it is straightforward to correlate hybridization differences with sequence differences, particularly with reference to the hybridization pattern of a known standard; and (iv) the ability to address each probe independently during synthesis, using high resolution photolithography, allows the array to be designed and optimized for any sequence. For example the length of any probe can be varied independently of the others.

The present tiling strategies result in sequencing and comparison methods suitable for routine large-scale practice with a high degree of confidence in the sequence output.

I. GENERAL TILING STRATEGIES

A. Selection of Reference Sequence

The chips are designed to contain probes exhibiting complementarity to one or more selected reference sequence whose sequence is known. The chips are used to read a target sequence comprising either the reference sequence itself or variants of that sequence. Target sequences may differ from the reference sequence at one or more positions but show a high overall degree of sequence identity with the reference sequence (*e.g.*, at least 75, 90, 95, 99, 99.9 or 99.99%). Any polynucleotide of known sequence can be selected as a reference sequence. Reference sequences of interest include sequences known to include mutations or polymorphisms associated with phenotypic changes having clinical significance in human patients. For example, the CFTR gene and P53 gene in humans have been identified as the location of several mutations resulting in cystic fibrosis or cancer respectively. Other reference sequences of interest include those that serve to identify pathogenic microorganisms and/or are the site of mutations by which such microorganisms acquire drug resistance (*e.g.*, the HIV reverse transcriptase gene). Other reference sequences of interest include regions where polymorphic variations are known to occur (*e.g.*, the D-loop region of mitochondrial DNA). These reference sequences have utility for, *e.g.*, forensic or epidemiological studies. Other reference sequences of interest include p34 (related to p53), p65 (implicated in breast, prostate and liver cancer), and DNA segments encoding cytochromes P450 (*see* Meyer et al., *Pharmac. Ther.* 46, 349-355 (1990)). Other reference sequences of interest include those from the genome of pathogenic viruses (*e.g.*, hepatitis (A, B, or C), herpes virus (*e.g.*, VZV, HSV-1, HAV-6, HSV-II, and CMV, Epstein Barr virus), adenovirus, influenza virus, flaviviruses, echovirus, rhinovirus, coxsackie virus, cornovirus, respiratory syncytial virus,

mumps virus, rotavirus, measles virus, rubella virus, parvovirus, vaccinia virus, HTLV virus, dengue virus, papillomavirus, molluscum virus, poliovirus, rabies virus, JC virus and arboviral encephalitis virus. Other reference sequences of interest are from genomes or episomes of pathogenic bacteria, particularly regions that confer drug resistance or allow phylogenic characterization of the host (*e.g.*, 16S rRNA or corresponding DNA). For example, such bacteria include chlamydia, rickettsial bacteria, mycobacteria, staphylococci, treptocci, pneumonococci, meningococci and conococci, klebsiella, proteus, serratia, pseudomonas, legionella, diphtheria, salmonella, bacilli, cholera, tetanus, botulism, anthrax, plague, leptospirosis, and Lymes disease bacteria. Other reference sequences of interest include those in which mutations result in the following autosomal recessive disorders: sickle cell anemia, β-thalassemia, phenylketonuria, galactosemia, Wilson's disease, hemochromatosis, severe combined immunodeficiency, alpha-1-antitrypsin deficiency, albinism, alkaptonuria, lysosomal storage diseases and Ehlers-Danlos syndrome. Other reference sequences of interest include those in which mutations result in X-linked recessive disorders: hemophilia, glucose-6-phosphate dehydrogenase, agammaglobulimenia, diabetes insipidus, Lesch-Nyhan syndrome, muscular dystrophy, Wiskott-Aldrich syndrome, Fabry's disease and fragile X-syndrome. Other reference sequences of interest includes those in which mutations result in the following autosomal dominant disorders: familial hypercholesterolemia, polycystic kidney disease, Huntingdon's disease, hereditary spherocytosis, Marfan's syndrome, von Willebrand's disease, neurofibromatosis, tuberous sclerosis, hereditary hemorrhagic telangiectasia, familial colonic polyposis, Ehlers-Danlos syndrome, myotonic dystrophy, muscular dystrophy, osteogenesis imperfecta, acute intermittent porphyria, and von Hippel-Lindau disease.

The length of a reference sequence can vary widely from a full-length genome, to an individual chromosome, episome, gene, component of a gene, such as an exon, intron or

regulatory sequences, to a few nucleotides. A reference sequence of between about 2, 5, 10, 20, 50, 100, 5000, 1000, 5,000 or 10,000, 20,000 or 100,000 nucleotides is common. Sometimes only particular regions of a sequence (*e.g.*, exons of a gene) are of interest. In such situations, the particular regions can be considered as separate reference sequences or can be considered as components of a single reference sequence, as matter of arbitrary choice.

A reference sequence can be any naturally occurring, mutant, consensus or purely hypothetical sequence of nucleotides, RNA or DNA. For example, sequences can be obtained from computer data bases, publications or can be determined or conceived *de novo*. Usually, a reference sequence is selected to show a high degree of sequence identity to envisaged target sequences. Often, particularly, where a significant degree of divergence is anticipated between target sequences, more than one reference sequence is selected. Combinations of wildtype and mutant reference sequences are employed in several applications of the tiling strategy.

B. Chip Design

1. Basic Tiling Strategy

The basic tiling strategy provides an array of immobilized probes for analysis of target sequences showing a high degree of sequence identity to one or more selected reference sequences. The strategy is first illustrated for an array that is subdivided into four probe sets, although it will be apparent that in some situations, satisfactory results are obtained from only two probe sets. A first probe set comprises a plurality of probes exhibiting perfect complementarity with a selected reference sequence. The perfect complementarity usually exists throughout the length of the probe. However, probes having a segment or segments of perfect complementarity that is/are flanked by leading or trailing sequences lacking complementarity to the reference sequence can also be used. Within a segment of complementarity, each probe in the first probe set has at

least one interrogation position that corresponds to a nucleotide in the reference sequence. That is, the interrogation position is aligned with the corresponding nucleotide in the reference sequence, when the probe and reference sequence are aligned to maximize complementarity between the two. If a probe has more than one interrogation position, each corresponds with a respective nucleotide in the reference sequence. The identity of an interrogation position and corresponding nucleotide in a particular probe in the first probe set cannot be determined simply by inspection of the probe in the first set. As will become apparent, an interrogation position and corresponding nucleotide is defined by the comparative structures of probes in the first probe set and corresponding probes from additional probe sets.

In principle, a probe could have an interrogation position at each position in the segment complementary to the reference sequence. Sometimes, interrogation positions provide more accurate data when located away from the ends of a segment of complementarity. Thus, typically a probe having a segment of complementarity of length x does not contain more than x-2 interrogation positions. Since probes are typically 9-21 nucleotides, and usually all of a probe is complementary, a probe typically has 1-19 interrogation positions. Often the probes contain a single interrogation position, at or near the center of probe.

For each probe in the first set, there are, for purposes of the present illustration, three corresponding probes from three additional probe sets. *See* Fig. 1. Thus, there are four probes corresponding to each nucleotide of interest in the reference sequence. Each of the four corresponding probes has an interrogation position aligned with that nucleotide of interest. Usually, the probes from the three additional probe sets are identical to the corresponding probe from the first probe set with one exception. The exception is that at least one (and often only one) interrogation position, which occurs in the same position in each of the four corresponding probes from the four probe sets, is occupied by a different nucleotide in the four probe sets. For example, for an A

nucleotide in the reference sequence, the corresponding probe from the first probe set has its interrogation position occupied by a T, and the corresponding probes from the additional three probe sets have their respective interrogation positions occupied by A, C, or G, a different nucleotide in each probe. Of course, if a probe from the first probe set comprises trailing or flanking sequences lacking complementarity to the reference sequences (*see* Fig. 2), these sequences need not be present in corresponding probes from the three additional sets. Likewise corresponding probes from the three additional sets can contain leading or trailing sequences outside the segment of complementarity that are not present in the corresponding probe from the first probe set. Occasionally, the probes from the additional three probe set are identical (with the exception of interrogation position(s)) to a contiguous subsequence of the full complementary segment of the corresponding probe from the first probe set. In this case, the subsequence includes the interrogation position and usually differs from the full-length probe only in the omission of one or both terminal nucleotides from the termini of a segment of complementarity. That is, if a probe from the first probe set has a segment of complementarity of length n, corresponding probes from the other sets will usually include a subsequence of the segment of at least length n-2. Thus, the subsequence is usually at least 3, 4, 7, 9, 15, 21, or 25 nucleotides long, most typically, in the range of 9-21 nucleotides. The subsequence should be sufficiently long to allow a probe to hybridize detectably more strongly to a variant of the reference sequence mutated at the interrogation position than to the reference sequence.

The probes can be oligodeoxyribonucleotides or oligoribonucleotides, or any modified forms of these polymers that are capable of hybridizing with a target nucleic sequence by complementary base-pairing. Complementary base pairing means sequence-specific base pairing which includes e.g., Watson-Crick base pairing as well as other forms of base pairing such as Hoogsteen base pairing. Modified forms

include 2'-O-methyl oligoribonucleotides and so-called PNAs, in which oligodeoxyribonucleotides are linked via peptide bonds rather than phophodiester bonds. The probes can be attached by any linkage to a support (*e.g.*, 3', 5' or via the base). 3' attachment is more usual as this orientation is compatible with the preferred chemistry for solid phase synthesis of oligonucleotides.

The number of probes in the first probe set (and as a consequence the number of probes in additional probe sets) depends on the length of the reference sequence, the number of nucleotides of interest in the reference sequence and the number of interrogation positions per probe. In general, each nucleotide of interest in the reference sequence requires the same interrogation position in the four sets of probes. Consider, as an example, a reference sequence of 100 nucleotides, 50 of which are of interest, and probes each having a single interrogation position. In this situation, the first probe set requires fifty probes, each having one interrogation position corresponding to a nucleotide of interest in the reference sequence. The second, third and fourth probe sets each have a corresponding probe for each probe in the first probe set, and so each also contains a total of fifty probes. The identity of each nucleotide of interest in the reference sequence is determined by comparing the relative hybridization signals at four probes having interrogation positions corresponding to that nucleotide from the four probe sets.

In some reference sequences, every nucleotide is of interest. In other reference sequences, only certain portions in which variants (*e.g.*, mutations or polymorphisms) are concentrated are of interest. In other reference sequences, only particular mutations or polymorphisms and immediately adjacent nucleotides are of interest. Usually, the first probe set has interrogation positions selected to correspond to at least a nucleotide (*e.g.*, representing a point mutation) and one immediately adjacent nucleotide. Usually, the probes in the first set have interrogation positions corresponding to at least 3, 10, 50, 100, 1000, or 20,000 contiguous

nucleotides. The probes usually have interrogation positions corresponding to at least 5, 10, 30, 50, 75, 90, 99 or sometimes 100% of the nucleotides in a reference sequence. Frequently, the probes in the first probe set completely span the reference sequence and overlap with one another relative to the reference sequence. For example, in one common arrangement each probe in the first probe set differs from another probe in that set by the omission of a 3' base complementary to the reference sequence and the acquisition of a 5' base complementary to the reference sequence. *See* Fig. 3.

For conceptual simplicity, the probes in a set are usually arranged in order of the sequence in a lane across the chip. A lane contains a series of overlapping probes, which represent or tile across, the selected reference sequence (*see* Fig. 3). The components of the four sets of probes are usually laid down in four parallel lanes, collectively constituting a row in the horizontal direction and a series of 4-member columns in the vertical direction. Corresponding probes from the four probe sets (*i.e.*, complementary to the same subsequence of the reference sequence) occupy a column. Each probe in a lane usually differs from its predecessor in the lane by the omission of a base at one end and the inclusion of additional base at the other end as shown in Fig. 3. However, this orderly progression of probes can be interrupted by the inclusion of control probes or omission of probes in certain columns of the array. Such columns serve as controls to orient the chip, or gauge the background, which can include target sequence nonspecifically bound to the chip.

The probes sets are usually laid down in lanes such that all probes having an interrogation position occupied by an A form an A-lane, all probes having an interrogation position occupied by a C form a C-lane, all probes having an interrogation position occupied by a G form a G-lane, and all probes having an interrogation position occupied by a T (or U) form a T lane (or a U lane). Note that in this arrangement there is not a unique correspondence between probe sets and lanes. Thus, the probe from the first probe set is laid down

in the A-lane, C-lane, A-lane, A-lane and T-lane for the five columns in Fig. 4. The interrogation position on a column of probes corresponds to the position in the target sequence whose identity is determined from analysis of hybridization to the probes in that column. Thus, I_1-I_5 respectively correspond to N_1-N_5 in Fig. 4. The interrogation position can be anywhere in a probe but is usually at or near the central position of the probe to maximize differential hybridization signals between a perfect match and a single-base mismatch. For example, for an 11 mer probe, the central position is the sixth nucleotide.

Although the array of probes is usually laid down in rows and columns as described above, such a physical arrangement of probes on the chip is not essential. Provided that the spatial location of each probe in an array is known, the data from the probes can be collected and processed to yield the sequence of a target irrespective of the physical arrangement of the probes on a chip. In processing the data, the hybridization signals from the respective probes can be reassorted into any conceptual array desired for subsequent data reduction whatever the physical arrangement of probes on the chip.

A range of lengths of probes can be employed in the chips. As noted above, a probe may consist exclusively of a complementary segments, or may have one or more complementary segments juxtaposed by flanking, trailing and/or intervening segments. In the latter situation, the total length of complementary segment(s) is more important that the length of the probe. In functional terms, the complementarity segment(s) of the first probe sets should be sufficiently long to allow the probe to hybridize detectably more strongly to a reference sequence compared with a variant of the reference including a single base mutation at the nucleotide corresponding to the interrogation position of the probe. Similarly, the complementarity segment(s) in corresponding probes from additional probe sets should be sufficiently long to allow a probe to hybridize detectably more strongly to a variant of the reference sequence having a single nucleotide

substitution at the interrogation position relative to the reference sequence. A probe usually has a single complementary segment having a length of at least 3 nucleotides, and more usually at least 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 or 30 bases exhibiting perfect complementarity (other than possibly at the interrogation position(s) depending on the probe set) to the reference sequence. In bridging strategies, where more than one segment of complementarity is present, each segment provides at least three complementary nucleotides to the reference sequence and the combined segments provide at least two segments of three or a total of six complementary nucleotides. As in the other strategies, the combined length of complementary segments is typically from 6-30 nucleotides, and preferably from about 9-21 nucleotides. The two segments are often approximately the same length. Often, the probes (or segment of complementarity within probes) have an odd number of bases, so that an interrogation position can occur in the exact center of the probe.

In some chips, all probes are the same length. Other chips employ different groups of probe sets, in which case the probes are of the same size within a group, but differ between different groups. For example, some chips have one group comprising four sets of probes as described above in which all the probes are 11 mers, together with a second group comprising four sets of probes in which all of the probes are 13 mers. Of course, additional groups of probes can be added. Thus, some chips contain, *e.g.*, four groups of probes having sizes of 11 mers, 13 mers, 15 mers and 17 mers. Other chips have different size probes within the same group of four probe sets. In these chips, the probes in the first set can vary in length independently of each other. Probes in the other sets are usually the same length as the probe occupying the same column from the first set. However, occasionally different lengths of probes can be included at the same column position in the four lanes. The different length probes are included to equalize hybridization signals from probes irrespective of

whether A-T or C-G bonds are formed at the interrogation position.

The length of probe can be important in distinguishing between a perfectly matched probe and probes showing a single-base mismatch with the target sequence. The discrimination is usually greater for short probes. Shorter probes are usually also less susceptible to formation of secondary structures. However, the absolute amount of target sequence bound, and hence the signal, is greater for larger probes. The probe length representing the optimum compromise between these competing considerations may vary depending on *inter alia* the GC content of a particular region of the target DNA sequence, secondary structure, synthesis efficiency and cross-hybridization. In some regions of the target, depending on hybridization conditions, short probes (*e.g.*, 11 mers) may provide information that is inaccessible from longer probes (*e.g.*, 19 mers) and vice versa. Maximum sequence information can be read by including several groups of different sized probes on the chip as noted above. However, for many regions of the target sequence, such a strategy provides redundant information in that the same sequence is read multiple times from the different groups of probes. Equivalent information can be obtained from a single group of different sized probes in which the sizes are selected to maximize readable sequence at particular regions of the target sequence. The appropriate size of probes at different regions of the target sequence can be determined from, *e.g.*, Fig. 12, which compares the readability of different sized probes in different regions of a target. The strategy of customizing probe length within a single group of probe sets minimizes the total number of probes required to read a particular target sequence. This leaves ample capacity for the chip to include probes to other reference sequences.

The invention provides an optimization block which allows systematic variation of probe length and interrogation position to optimize the selection of probes for analyzing a particular nucleotide in a reference sequence. The block comprises alternating columns of probes complementary to the

wildtype target and probes complementary to a specific mutation. The interrogation position is varied between columns and probe length is varied down a column. Hybridization of the chip to the reference sequence or the mutant form of the reference sequence identifies the probe length and interrogation position providing the greatest differential hybridization signal.

The probes are designed to be complementary to either strand of the reference sequence (*e.g.*, coding or non-coding). Some chips contain separate groups of probes, one complementary to the coding strand, the other complementary to the noncoding strand. Independent analysis of coding and noncoding strands provides largely redundant information. However, the regions of ambiguity in reading the coding strand are not always the same as those in reading the noncoding strand. Thus, combination of the information from coding and noncoding strands increases the overall accuracy of sequencing.

Some chips contain additional probes or groups of probes designed to be complementary to a second reference sequence. The second reference sequence is often a subsequence of the first reference sequence bearing one or more commonly occurring mutations or interstrain variations. The second group of probes is designed by the same principles as described above except that the probes exhibit complementarity to the second reference sequence. The inclusion of a second group is particular useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (*i.e.*, two or more mutations within 9 to 21 bases). Of course, the same principle can be extended to provide chips containing groups of probes for any number of reference sequences. Alternatively, the chips may contain additional probe(s) that do not form part of a tiled array as noted above, but rather serves as probe(s) for a conventional reverse dot blot. For example, the presence of mutation can be detected from binding of a target sequence to a single oligomeric probe harboring the mutation. Preferably, an

additional probe containing the equivalent region of the wildtype sequence is included as a control.

The chips are read by comparing the intensities of labelled target bound to the probes in an array. Specifically, a comparison is performed between each lane of probes (*e.g.*, A, C, G and T lanes) at each columnar position (physical or conceptual). For a particular columnar position, the lane showing the greatest hybridization signal is called as the nucleotide present at the position in the target sequence corresponding to the interrogation position in the probes. *See* Fig. 5. The corresponding position in the target sequence is that aligned with the interrogation position in corresponding probes when the probes and target are aligned to maximize complementarity. Of the four probes in a column, only one can exhibit a perfect match to the target sequence whereas the others usually exhibit at least a one base pair mismatch. The probe exhibiting a perfect match usually produces a substantially greater hybridization signal than the other three probes in the column and is thereby easily identified. However, in some regions of the target sequence, the distinction between a perfect match and a one-base mismatch is less clear. Thus, a call ratio is established to define the ratio of signal from the best hybridizing probes to the second best hybridizing probe that must be exceeded for a particular target position to be read from the probes. A high call ratio ensures that few if any errors are made in calling target nucleotides, but can result in some nucleotides being scored as ambiguous, which could in fact be accurately read. A lower call ratio results in fewer ambiguous calls, but can result in more erroneous calls. It has been found that at a call ratio of 1.2 virtually all calls are accurate. However, a small but significant number of bases (*e.g.*, up to about 10%) may have to be scored as ambiguous.

Although small regions of the target sequence can sometimes be ambiguous, these regions usually occur at the same or similar segments in different target sequences. Thus, for precharacterized mutations, it is known in advance whether

that mutation is likely to occur within a region of unambiguously determinable sequence.

An array of probes is most useful for analyzing the reference sequence from which the probes were designed and variants of that sequence exhibiting substantial sequence similarity with the reference sequence (*e.g.*, several single-base mutants spaced over the reference sequence). When an array is used to analyze the exact reference sequence from which it was designed, one probe exhibits a perfect match to the reference sequence, and the other three probes in the same column exhibits single-base mismatches. Thus, discrimination between hybridization signals is usually high and accurate sequence is obtained. High accuracy is also obtained when an array is used for analyzing a target sequence comprising a variant of the reference sequence that has a single mutation relative to the reference sequence, or several widely spaced mutations relative to the reference sequence. At different mutant loci, one probe exhibits a perfect match to the target, and the other three probes occupying the same column exhibit single-base mismatches, the difference (with respect to analysis of the reference sequence) being the lane in which the perfect match occurs.

For target sequences showing a high degree of divergence from the reference strain or incorporating several closely spaced mutations from the reference strain, a single group of probes (*i.e.*, designed with respect to a single reference sequence) will not always provide accurate sequence for the highly variant region of this sequence. At some particular columnar positions, it may be that no single probe exhibits perfect complementarity to the target and that any comparison must be based on different degrees of mismatch between the four probes. Such a comparison does not always allow the target nucleotide corresponding to that columnar position to be called. Deletions in target sequences can be detected by loss of signal from probes having interrogation positions encompassed by the deletion. However, signal may also be lost from probes having interrogation positions closely proximal to the deletion resulting in some regions of the target sequence

that cannot be read. Target sequence bearing insertions will also exhibit short regions including and proximal to the insertion that usually cannot be read.

The presence of short regions of difficult-to-read target because of closely spaced mutations, insertions or deletion, does not prevent determination of the remaining sequence of the target as different regions of a target sequence are determined independently. Moreover, such ambiguities as might result from analysis of diverse variants with a single group of probes can be avoided by including multiple groups of probe sets on a chip. For example, one group of probes can be designed based on a full-length reference sequence, and the other groups on subsequences of the reference sequence incorporating frequently occurring mutations or strain variations.

A particular advantage of the present sequencing strategy over conventional sequencing methods is the capacity simultaneously to detect and quantify proportions of multiple target sequences. Such capacity is valuable, *e.g.*, for diagnosis of patients who are heterozygous with respect to a gene or who are infected with a virus, such as HIV, which is usually present in several polymorphic forms. Such capacity is also useful in analyzing targets from biopsies of tumor cells and surrounding tissues. The presence of multiple target sequences is detected from the relative signals of the four probes at the array columns corresponding to the target nucleotides at which diversity occurs. The relative signals at the four probes for the mixture under test are compared with the corresponding signals from a homogeneous reference sequence. An increase in a signal from a probe that is mismatched with respect to the reference sequence, and a corresponding decrease in the signal from the probe which is matched with the reference sequence signal the presence of a mutant strain in the mixture. The extent in shift in hybridization signals of the probes is related to the proportion of a target sequence in the mixture. Shifts in relative hybridization signals can be quantitatively related to proportions of reference and mutant sequence by prior

calibration of the chip with seeded mixtures of the mutant and reference sequences. By this means, a chip can be used to detect variant or mutant strains constituting as little as 1, 5, 20, or 25 % of a mixture of stains.

Similar principles allow the simultaneous analysis of multiple target sequences even when none is identical to the reference sequence. For example, with a mixture of two target sequences bearing first and second mutations, there would be a variation in the hybridization patterns of probes having interrogation positions corresponding to the first and second mutations relative to the hybridization pattern with the reference sequence. At each position, one of the probes having a mismatched interrogation position relative to the reference sequence would show an increase in hybridization signal, and the probe having a matched interrogation position relative to the reference sequence would show a decrease in hybridization signal. Analysis of the hybridization pattern of the mixture of mutant target sequences, preferably in comparison with the hybridization pattern of the reference sequence, indicates the presence of two mutant target sequences, the position and nature of the mutation in each strain, and the relative proportions of each strain.

In a variation of the above method, the different components in a mixture of target sequences are differentially labelled before being applied to the array. For example, a variety of fluorescent labels emitting at different wavelength are available. The use of differential labels allows independent analysis of different targets bound simultaneously to the array. For example, the methods permit comparison of target sequences obtained from a patient at different stages of a disease.

2. Omission of Probes

The general strategy outlined above employs four probes to read each nucleotide of interest in a target sequence. One probe (from the first probe set) shows a perfect match to the reference sequence and the other three probes (from the second, third and fourth probe sets) exhibit a mismatch with

the reference sequence and a perfect match with a target sequence bearing a mutation at the nucleotide of interest. The provision of three probes from the second, third and fourth probe sets allows detection of each of the three possible nucleotide substitutions of any nucleotide of interest. However, in some reference sequences or regions of reference sequences, it is known in advance that only certain mutations are likely to occur. Thus, for example, at one site it might be known that an A nucleotide in the reference sequence may exist as a T mutant in some target sequences but is unlikely to exist as a C or G mutant. Accordingly, for analysis of this region of the reference sequence, one might include only the first and second probe sets, the first probe set exhibiting perfect complementarity to the reference sequence, and the second probe set having an interrogation position occupied by an invariant A residue (for detecting the T mutant). In other situations, one might include the first, second and third probes sets (but not the fourth) for detection of a wildtype nucleotide in the reference sequence and two mutant variants thereof in target sequences. In some chips, probes that would detect silent mutations (*i.e.*, not affecting amino acid sequence) are omitted.

In some chips, the probes from the first probe set are omitted corresponding to some or all positions of the reference sequences. Such chips comprise at least two probe sets. The first probe set has a plurality of probes. Each probe comprises a segment exactly complementary to a subsequence of a reference sequence except in at least one interrogation position. A second probe set has a corresponding probe for each probe in the first probe set. The corresponding probe in the second probe set is identical to a sequence comprising the corresponding probe form the first probe set or a subsequence thereof that includes the at least one (and usually only one) interrogation position except that the at least one interrogation position is occupied by a different nucleotide in each of the two corresponding probes from the first and second probe sets. A third probe set, if present, also comprises a corresponding probe for each probe

in the first probe set except at the at least one interrogation position, which differs in the corresponding probes from the three sets. Omission of probes having a segment exhibiting perfect complementarity to the reference sequence results in loss of control information, *i.e.*, the detection of nucleotides in a target sequence that are the same as those in a reference sequence. However, similar information can be obtained by hybridizing a chip lacking probes from the first probe set to both target and reference sequences. The hybridization can be performed sequentially, or concurrently, if the target and reference are differentially labelled. In this situation, the presence of a mutation is detected by a shift in the background hybridization intensity of the reference sequence to a perfectly matched hybridization signal of the target sequence, rather than by a comparison of the hybridization intensities of probes from the first set with corresponding probes from the second, third and fourth sets.

3. Wildtype Probe Lane

When the chips comprise four probe sets, as discussed *supra*, and the probe sets are laid down in four lanes, an A lane, a C-lane, a G lane and a T or U lane, the probe having a segment exhibiting perfect complementarity to a reference sequence varies between the four lanes from one column to another. This does not present any significant difficulty in computer analysis of the data from the chip. However, visual inspection of the hybridization pattern of the chip is sometimes facilitated by provision of an extra lane of probes, in which each probe has a segment exhibiting perfect complementarity to the reference sequence. *See* Fig. 4. This segment is identical to a segment from one of the probes in the other four lanes (which lane depending on the column position). The extra lane of probes (designated the wildtype lane) hybridizes to a target sequence at all nucleotide positions except those in which deviations from the reference sequence occurs. The hybridization pattern of the wildtype lane thereby provides a simple visual indication of mutations.

4. Deletion, Insertion and Multiple-Mutation Probes

Some chips provide an additional probe set specifically designed for analyzing deletion mutations. The additional probe set comprises a probe corresponding to each probe in the first probe set as described above. However, a probe from the additional probe set differs from the corresponding probe in the first probe set in that the nucleotide occupying the interrogation position is deleted in the probe from the additional probe set. *See* Fig. 6. Optionally, the probe from the additional probe set bears an additional nucleotide at one of its termini relative to the corresponding probe from the first probe set. The probe from the additional probe set will hybridize more strongly than the corresponding probe from the first probe set to a target sequence having a single base deletion at the nucleotide corresponding to the interrogation position. Additional probe sets are provided in which not only the interrogation position, but also an adjacent nucleotide is detected.

Similarly, other chips provide additional probe sets for analyzing insertions. For example, one additional probe set has a probe corresponding to each probe in the first probe set as described above. However, the probe in the additional probe set has an extra T nucleotide inserted adjacent to the interrogation position. *See* Fig. 6. Optionally, the probe has one fewer nucleotide at one of its termini relative to the corresponding probe from the first probe set. The probe from the additional probe set hybridizes more strongly than the corresponding probe from the first probe set to a target sequence having an A nucleotide inserted in a position adjacent to that corresponding to the interrogation position. Similar additional probe sets are constructed having C, G or T/U nucleotides inserted adjacent to the interrogation position. Usually, four such probe sets, one for each nucleotide, are used in combination.

Other chips provide additional probes (multiple-mutation probes) for analyzing target sequences having multiple closely spaced mutations. A multiple-mutation probe is usually identical to a corresponding probe from the first set as

described above, except in the base occupying the interrogation position, and except at one or more additional positions, corresponding to nucleotides in which substitution may occur in the reference sequence. The one or more additional positions in the multiple mutation probe are occupied by nucleotides complementary to the nucleotides occupying corresponding positions in the reference sequence when the possible substitutions have occurred.

5. Block Tiling

As noted in the discussion of the general tiling strategy, a probe in the first probe set sometimes has more than one interrogation position. In this situation, a probe in the first probe set is sometimes matched with multiple groups of at least one, and usually, three additional probe sets. *See* Fig. 7. Three additional probe sets are used to allow detection of the three possible nucleotide substitutions at any one position. If only certain types of substitution are likely to occur (*e.g.*, transitions), only one or two additional probe sets are required (analogous to the use of probes in the basic tiling strategy). To illustrate for the situation where a group comprises three additional probe sets, a first such group comprises second, third and fourth probe sets, each of which has a probe corresponding to each probe in the first probe set. The corresponding probes from the second, third and fourth probes sets differ from the corresponding probe in the first set at a first of the interrogation positions. Thus, the relative hybridization signals from corresponding probes from the first, second, third and fourth probe sets indicate the identity of the nucleotide in a target sequence corresponding to the first interrogation position. A second group of three probe sets (designated fifth, sixth and seventh probe sets), each also have a probe corresponding to each probe in the first probe set. These corresponding probes differ from that in the first probe set at a second interrogation position. The relative hybridization signals from corresponding probes from the first, fifth, sixth, and seventh probe sets indicate the identity of the nucleotide in the target sequence

corresponding to the second interrogation position. As noted above, the probes in the first probe set often have seven or more interrogation positions. If there are seven interrogation positions, there are seven groups of three additional probe sets, each group of three probe sets serving to identify the nucleotide corresponding to one of the seven interrogation positions.

Each block of probes allows short regions of a target sequence to be read. For example, for a block of probes having seven interrogation positions, seven nucleotides in the target sequence can be read. Of course, a chip can contain any number of blocks depending on how many nucleotides of the target are of interest. The hybridization signals for each block can be analyzed independently of any other block. The block tiling strategy can also be combined with other tiling strategies, with different parts of the same reference sequence being tiled by different strategies.

The block tiling strategy offers two advantages over the basic strategy in which each probe in the first set has a single interrogation position. One advantage is that the same sequence information can be obtained from fewer probes. A second advantage is that each of the probes constituting a block (*i.e.*, a probe from the first probe set and a corresponding probe from each of the other probe sets) can have identical 3' and 5' sequences, with the variation confined to a central segment containing the interrogation positions. The identity of 3' sequence between different probes simplifies the strategy for solid phase synthesis of the probes on the chip and results in more uniform deposition of the different probes on the chip, thereby in turn increasing the uniformity of signal to noise ratio for different regions of the chip. A third advantage is that greater signal uniformity is achieved within a block.

6. Multiplex Tiling

In the block tiling strategy discussed above, the identity of a nucleotide in a target or reference sequence is determined by comparison of hybridization patterns of one

probe having a segment showing a perfect match with that of other probes (usually three other probes) showing a single base mismatch. In multiplex tiling, the identity of at least two nucleotides in a reference or target sequence is determined by comparison of hybridization signal intensities of four probes, two of which have a segment showing perfect complementarity or a single base mismatch to the reference sequence, and two of which have a segment showing perfect complementarity or a double-base mismatch to a segment. The four probes whose hybridization patterns are to be compared each have a segment that is exactly complementary to a reference sequence except at two interrogation positions, in which the segment may or may not be complementary to the reference sequence. The interrogation positions correspond to the nucleotides in a reference or target sequence which are determined by the comparison of intensities. The nucleotides occupying the interrogation positions in the four probes are selected according to the following rule. The first interrogation position is occupied by a different nucleotide in each of the four probes. The second interrogation position is also occupied by a different nucleotide in each of the four probes. In two of the four probes, designated the first and second probes, the segment is exactly complementary to the reference sequence except at not more than one of the two interrogation positions. In other words, one of the interrogation positions is occupied by a nucleotide that is complementary to the corresponding nucleotide from the reference sequence and the other interrogation position may or may not be so occupied. In the other two of the four probes, designated the third and fourth probes, the segment is exactly complementary to the reference sequence except that both interrogation positions are occupied by nucleotides which are noncomplementary to the respective corresponding nucleotides in the reference sequence.

There are number of ways of satisfying these conditions depending on whether the two nucleotides in the reference sequence corresponding to the two interrogation positions are the same or different. If these two nucleotides are different

in the reference sequence (probability 3/4), the conditions are satisfied by each of the two interrogation positions being occupied by the same nucleotide in any given probe. For example, in the first probe, the two interrogation positions would both be A, in the second probe, both would be C, in the third probe, each would be G, and in the fourth probe each would be T or U. If the two nucleotides in the reference sequence corresponding to the two interrogation positions are different, the conditions noted above are satisfied by each of the interrogation positions in any one of the four probes being occupied by complementary nucleotides. For example, in the first probe, the interrogation positions could be occupied by A and T, in the second probe by C and G, in the third probe by G and C, and in the four probe, by T and A. *See* (Fig. 8).

When the four probes are hybridized to a target that is the same as the reference sequence or differs from the reference sequence at one (but not both) of the interrogation positions, two of the four probes show a double-mismatch with the target and two probes show a single mismatch. The identity of probes showing these different degrees of mismatch can be determined from the different hybridization signals. From the identity of the probes showing the different degrees of mismatch, the nucleotides occupying both of the interrogation positions in the target sequence can be deduced.

For ease of illustration, the multiplex strategy has been initially described for the situation where there are two nucleotides of interest in a reference sequence and only four probes in an array. Of course, the strategy can be extended to analyze any number of nucleotides in a target sequence by using additional probes. In one variation, each pair of interrogation positions is read from a unique group of four probes. In a block variation, different groups of four probes exhibit the same segment of complementarity with the reference sequence, but the interrogation positions move within a block. The block and standard multiplex tiling variants can of course be used in combination for different regions of a reference sequence. Either or both variants can also be used in combination with any of the other tiling strategies described.

7. Helper Mutations

Occasionally small regions of a reference sequence give a low hybridization signal as a result of annealing of probes. The self-annealing reduces the amount of probe effectively available for hybridizing to the target. Although such regions of the target are generally small and the reduction of hybridization signal is usually not so substantial as to obscure the sequence of this region, this concern can be avoided by the use of probes incorporating helper mutations. The helper mutation(s) serve to break-up regions of internal complementarity within a probe and thereby prevent annealing. Usually, one or two helper mutations are quite sufficient for this purpose. The inclusion of helper mutations can be beneficial in any of the tiling strategies noted above. In general each probe having a particular interrogation position has the same helper mutation(s). Thus, such probes have a segment in common which shows perfect complementarity with a reference sequence, except that the segment contains at least one helper mutation (the same in each of the probes) and at least one interrogation position (different in all of the probes). For example, in the basic tiling strategy, a probe from the first probe set comprises a segment containing an interrogation position and showing perfect complementarity with a reference sequence except for one or two helper mutations. The corresponding probes from the second, third and fourth probe sets usually comprise the same segment (or sometimes a subsequence thereof including the helper